78945-30

- 1 -

LABEL SWITCHED ROUTING SYSTEM AND METHOD

Field of the Invention

The invention relates to systems and methods for performing label switched routing.

5 ## Background of the Invention

If there are multiple LSPs (label switched paths) that all originate on one LSR (label switched router) and all terminate on another LSR, then these LSPs can be merged (under control of the head-end LSR) into a single FA-LSP (forwarding

10 adjacency-label switched path) using the concept of link bundling which is described in detail in draft-kompella-mpls-bundle (see for example www.ietf.org/internet-drafts/draft-kompella-mpls-bundle-05.txt).

For example, to improve scalability of MPLS-TE

15 (multiple protocol label switching protocol - traffic engineering) it may be useful to aggregate TE LSPs. The aggregation is accomplished by:

an LSR creating a TE LSP;

the LSR forming a forwarding adjacency out of that

20 LSP (advertising this LSP as a link into an internal routing protocol such as ISIS or OSPF);

allowing other LSRs to use forwarding adjacencies for their path computation; and

nesting of LSPs originated by other LSRs into that

25 LSP by using a label stack construct.

The details of this approach and the label stack constructs can be found in draft-ietf-mpls-lsp-hierarchy, - (see for example www.ietf.org/internet-drafts/draft-ietf-mpls-lsp-hierarchy-02.txt).

30 This approach will be described further by way of example with reference to Figure 1. Shown is a network of

hierarchically connected nodes including at a lowest level in a hierarchy four nodes 1,2,10,11, at a higher level in the hierarchy four nodes 3,4,8,9, and at a highest level in the hierarchy three nodes 5,6,7. Also shown are four end user

5   terminal devices T1, T2, T3 and T4 connected to nodes 1,10,2,11 respectively. For a terminal T1 connected to node 1 to communicate with a terminal T2 connected to node 10, use may be made of a first forwarding adjacency established between node 5 and 7, a second between nodes 3 and 8, and a third between

10 nodes 1 and 10. The result is that a user packet to be forwarded from T1 to T2, say an IP (Internet Protocol) packet, will have the user packet header, a first label understood by nodes 1 and 10 representing the LSP between these nodes, a second label understood by nodes 3 and 8 representing the

15 forwarding adjacency between these nodes, and a third label understood by nodes 5 and 7 representing the forwarding adjacency between those nodes. Thus, for each packet there is the original user header, an IP header in this example, plus three labels. For packets which are long, the overhead

20 introduced by these three additional labels may not be significant. However, for short packets, the overhead can be a significant percentage of the overall packet size. In some networks, for example networks in which there is significant voice traffic, there is a high percentage of the overall packet

25 flow which has short packet length. An example packet size distribution is shown in Figure 2 obtained during a seven minute interval over a real network in March of 1998 (see www.caida.org/outreach/resources/learn/packetsizes) where it can be seen that a significant fraction of the packets have a

30 length less than 100 bytes. It is noted that in today's applications the number of small packets would be even larger than that shown in Figure 2 because the number the voice-over-IP and IP teleconference applications has increased. Using the above described hierarchical labeling approach in a network

with this type of packet length distribution would result in a
significant increase in overall overhead.

Summary of the Invention

5   A broad aspect of the invention provides a packet
routing/switching method. The method involves defining a
hierarchical plurality of label switched paths/forwarding
adjacency-label switched paths (LSP/FA-LSP) through a network
of nodes from a lowest (least-nested) level to a highest (most-

10   nested) level, each LSP/FA-LSP comprising a respective sequence
of nodes comprising at least a source node, a destination node,
and possibly one or more transit nodes. To route a packet flow
from a first source node of the network of nodes to a first
destination node of the network of nodes the following steps

15   are performed:

          a) maintaining at the first node a mapping between
the packet flow and a first LSP of the hierarchical plurality
of LSPs defined between the first source node and the first
destination node;

20   b) at the first source node, for each packet of said
packet flow, adding to the packet label switched routing
information comprising an LSP label identifying the first LSP
and sending the packet to subsequent node(s) in the sequence of
nodes defined for the first LSP;

25   c) at each node to which the packet is
routed/switched other than said first source node:

                    i) if the node is a source node of a higher
level FA-LSP than the LSP/FA-LSP identified by the LSP label of
the packet, changing the LSP label in the label switched

30   routing information to indicate the source node of the higher
level FA-LSP, and including in the label switched routing
information hierarchy information in respect of all lower level

LSPs/FA-LSPs in the hierarchy leading up to the higher level FA-LSP and forwarding the packet to the next node in the sequence of nodes defined for the higher level FA-LSP;

ii) if the node is only a transit node,
5   forwarding the packet to the next node in the sequence of nodes defined for the LSP/FA-LSP identified by the LSP label;

iii) if the node is a destination node of a higher level FA-LSP, changing the LSP label in the label switched routing information to indicate the source node of the
10 next lower level LSP/FA-LSP indicated by the hierarchy information, and changing the hierarchy information to include only hierarchy information in respect of LSPs/FA-LSPs in the hierarchy leading up to but not including the next lower level LSP/FA-LSP, and forwarding the packet to the next node in the
15 sequence of nodes defined for the next lower level LSP/FA-LSP.

Advantageously, the hierarchy information included in the packets takes significantly less space than the space required to include a full LSP label for each level in the hierarchy.

20   Preferably, for at least one of the LSPs/FA-LSPs in the hierarchical plurality of LSPs/FA-LSPs, an associated restoration path is defined between the source node and the destination node. Then, in each packet being routed/switched according to one of those LSPs/FA-LSPs an indication is
25 included of whether the packet should be routed/switched on the restoration path associated with the LSP/FA-LSP or not. Advantageously this allows for quick switching between normal and restoration paths by simply changing the indication.

Preferably, to allow routing/switching based on the
30 information added to the packets in the above manner, each node maintains information in association with every defined LSP/FA-LSP. The information for each defined LSP/FA-LSP has an LSP

label used to uniquely identify the LSP/FA-LSP throughout the network, an identification of the respective sequence of nodes, and an identification of the LSP/FA-LSP label for each possible next lowest level LSP/FA-LSP inside which the defined LSP/FA-
5 LSP may be nested.

Furthermore, for each packet, the hierarchy information preferably includes a component identifier associated with each level in the hierarchy below the level of the LSP label of the packet. The component associated with one
10 level below the level of the LSP label of the packet, when present, allows an identification of a particular possible next lowest level LSP/FA-LSP inside which the LSP/FA-LSP defined by the LSP label is to be used in routing the packet. The components associated with subsequent lower levels allow an
15 identification of a particular nested hierarchy of LSPs/FA-LSPs to be used for the packet.

In the event restoration paths are being provided, the information maintained in association with defined LSPs/FA-LSPs further defines source node, transit node, destination
20 node identifiers for the restoration path when present.

Preferably, the hierarchy information includes a bit position for each possible component at each level in the hierarchy, with a particular bit position being set (or cleared) to indicate a selected component as the particular
25 possible component. More generally, a multi-bit component identifier may be employed for each level. The component identifier must contain enough bits to uniquely distinguish between possible components.

According to another broad aspect, the invention
30 provides a method to be executed at a node within a network of interconnected nodes within which a hierarchical plurality of LSPs/FA-LSPs has been defined of performing label switching of packets having an LSP label and having a possibly empty

78945-30

- 6 -

components label. The method involves the node maintaining, in
a table for example, for each LSP/FA-LSP an identification of a
source node, transit nodes if any, and a destination node, and
for each LSP/FA-LSP an identification of all possible next

5   lowest level LSPs/FA-LSPs which may use the LSP/FA-LSP. For
each packet received the node obtains the LSP label, the LSP
label defining a current LSP/FA-LSP. The node obtains the
components label. The node looks the information for the LSP
label. In the event the node is a source node of a next higher

10 level FA-LSP of which the current LSP/FA-LSP forms a component,
the node switches the LSP label to contain the label of the
next higher level FA-LSP which is used by the current LSP/FA-
LSP, and changes the components label to include in a first
component identifier an identifier of the current LSP/FA-LSP.

15 In the event the node is the destination node of the current
LSP/FA-LSP, the node changes the LSP label to the LSP/FA-LSP
label for the lower level hierarchy determined from the
components label and the table. The node re-applies the
components label, re-applies the LSP label, and changes an

20 output interface such that the packet is forwarded to an
appropriate next node.

        According to another broad aspect, the invention
provides a method of performing label switched routing. The
method involves, at each node in a network of nodes, for each

25 packet removing a previous LSP header and adding a new header
containing a full LSP label for a current LSP/FA-LSP, and
containing components identifiers which allow local
identification of a hierarchy of LSPs/FA-LSPs of which the
current LSP/FA-LSP forms a part.

30 Brief Description of the Drawings

        Preferred embodiments of the invention will now be
described with reference to the attached drawings in which:

Figure 1 is a block diagram of an example network in which label switched routing may employed;

Figure 2 is a plot of an example packet length distribution;

5 Figure 3 is a block diagram of an example network in which an embodiment of the invention may be employed;

Figure 4 is an example cell format provided by an embodiment of the invention;

Figures 5A through 5F are cell formats used between
10 adjacent nodes of Figure 3 in an example implementation of the invention;

Figures 6A through 6G are cell formats used between adjacent nodes of Figure 3 in an example implementation of the invention in which the restoration path is being employed;

15 Figure 7 is a flowchart of an example method of processing cells by each node in a network; and

Figure 8 is a table of information maintained for the LSP/FA-LSPs of Figure 3.

Detailed Description of the Preferred Embodiments

20 According to an embodiment of the invention, label switched routing is performed within a hierarchy of LSP/FA-LSPs defined/provisioned in a network of interconnected nodes. Rather than transmitting an entire label stack with an LSP label for each LSP/FA-LSP in the hierarchy, at a given node in
25 the network, a single LSP label is transmitted together with a components label which contains a list of component identifiers which do not inherently identify LSP/FA-LSPs, but from which the full LSP/FA-LSP labels can be determined locally at each node using previously distributed information described in
30 detail below. The list of component identifiers is a shorthand way identifying to adjacent nodes the identity of the LSP/FA-

- 8 -

LSP hierarchy without transmitting the entire LSP labels and thus significantly reducing overhead. The single label which is transmitted is the label of the LSP/FA-LSP of which the given node forms a part.

5   The hierarchy of LSP/FA-LSPs will be considered to contain a plurality of levels. The lowest level will involve LSPs defined between edge nodes in the network. These lowest level LSPs may employ other FA-LSPs in a nested manner, with each level of nesting representing a higher level in the
10   hierarchy. LSPs are provisioned at the lowest level in the hierarchy. An FA-LSP is a bundle of at least two LSPs.

          For a bundle at level k+1 in a hierarchy of levels, the single label will identify the LSP/FA-LSP at level k+1, and the list components will specify the components at levels 1
15   through k. These components will be unique in the network. For a bundle at level 1 in the hierarchy of levels, this being the lowest level, k+1 = 1, so k = 0. The single label will identify the LSP/FA-LSP at level k+1 = 1, and there will be no components since there are no additional components at this
20   level. For a bundle at level 2 in the hierarchy, the single label will identify the LSP/FA-LSP at level 2, and there will be a single component identifier for level 1 which will allow the identification of the level 1 LSP/FA-LSP label at nodes within the network, and so on.

25   The information about each LSP/FA-LSP provisioned within the network is made known at each node including the relationship between the components and actual LSP/FA-LSPs which may for example be maintained in a table on each node. Such a table can be built using any suitable means. For example
30   the table can be built "In band" meaning the data network (DN) used to transport the user data will also transport the information needed to build the table. The table could also be built "Out of band" meaning the information needed to build

the table is transported over a control network (CN) that is different from data network.

Having established this information at each node, a cell format is used between nodes which will include the LSP 5 label of the current LSP/FA-LSP, and a component for each lower level in the hierarchy. In a preferred embodiment, the possible components are maintained in a list in association with each LSP/FA-LSP, and the LSP identifier is simply used to identify the position in the list. For example, if there are 10 four possible components for a given LSP/FA-LSP, then these four components are identified in the table in association with the LSP/FA-LSP. Then, the component identifier is used to identify a position in the list of four components, and thereby identify one of four LSP/FA-LSPs. This might for example 15 involve allocating a single bit in the components label for each possible component, and then setting one of four bits high (or low) to indicate a selected component. In this manner, one of four pre-determined LSP labels can be indicated with only four bits in the header, this being significantly less than the 20 size of an entire LSP label.

This embodiment of the invention will now be described in further detail with reference to the example of Figure 3 which is a network containing 13 interconnected nodes 31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47. It is 25 assumed that at the lowest in the hierarchy, there is an LSP A provisioned between nodes 31 and 40 and an LSP B between nodes 32 and 41. At the next level in the hierarchy, there is an FA-LSP K provisioned between nodes 33 and 38, and an FA-LSP L provisioned between nodes 34 and 39. An FA-LSP C is 30 provisioned between nodes 44 and 46, and an FA-LSP D is provisioned between nodes 45 and 47. At the next level in the hierarchy (the highest level in this example), there is an FA-LSP X provisioned between nodes 35 and 37 which passes through

78945-30

- 10 -

node 36, and for which there is a provisioned a restoration
path which passes through nodes 42 and 43.

For this example, it is assumed that there are two
user packet streams entering nodes 31, 32 respectively from end
5   user terminals T5,T7 and exiting nodes 40,41 to end user
terminals T6,T8 respectively. These may contain any
appropriate type of user packet, for example IP packets.

In the nodes at the edge of the network (like 31, 32,
40, 41 there is maintained a table (or other suitable
10 structure) that defines a correspondence between the
destination and the LSP label for the lowest level in the
hierarchy. Thus, in our example, for the user packet stream
entering node 31 from end user device T4, node 31 maintains a
correspondence between the destination for device T6 and the
15 LSP label representing the path between nodes 31 and 40, namely
LSP A. Similarly, for the user packet stream entering node 32
from end user device T7, node 32 maintains a correspondence
between the destination for device T8 and the LSP label
representing the path between nodes 32 and 40, namely B. More
20 generally, some association with packet streams and LSPs needs
to be maintained.

In the nodes internal to the edge, a table (or other
suitable structure) is provided with information concerning
each LSP and FA-LSP. The information might for example take
25 the format of the table shown in Fig. 8, where data has been
filled in for the example network of Figure 3. Figure 8
contains a record for each LSP/FA-LSP and for the illustrated
example this results in a record for each of LSP/FA-LSPs A, B,
C, D, K, L and X. In the above example, the first column
30 contains the full LSP/FA-LSP label this typically takes 32
bits. The second column contains the source node for the
LSP/FA-LSP. The third column contains the transit nodes for
the LSP/FA-LSP and the fourth column contains the destination

78945-30

- 11 -

node for the LSP/FA-LSP. Node identifiers could be in any
suitable format for example integer, IP addresses or even a
character string. For LSP A in our example, there is source
node 31, transit nodes 33,35,36,37,38 and destination node 40.

5   In a preferred embodiment, additional information is
provided which allows an identification of restoration paths
for one or more LSP/FA-LSPs. For example, in Figure 8, the
fifth, sixth and seventh identify source, transit and
destination nodes for a restoration (backup) path should one
10 have been provisioned for the particular LSP/FA-LSP. For FA-
LSP X in our example, there is a provisioned backup path which
has source node 35, transit nodes 42, 43 and destination node
37. In the eighth through 11th columns, there are places for
"components". A component for this purpose is an LSP label of
15 an LSP/FA-LSP which is at the lower level in the LSP/FA-LSP
hierarchy than that of the current record under consideration,
and which may make use of the LSP/FA-LSP of the record under
consideration. For LSP A, this is not made use of by a lower
level LSP/FA-LSP and as such has no components listed. For
20 FA-LSP K, this may be used in LSP A, or in LSP B, and as such
there is space for the identity of these two components.
Similarly, for PA-LSP L, there are first and second components
C and D, and for FA-LSP X, there are first and second
components K and L.

25   This table can be also used to get the source for a
flow to which an error message (icmp for IP) is to be sent.
This table can be used in conjunction with an instance of a
control network (CN) with out of band signalization.

            Referring to Figure 4, an example cell format will be
30 described. The cell format has a current label field 60 which
is the basis of the current routing/switching process. For
embodiments allowing the specification of restoration paths,
there is a field 62 for indicating whether or not the

- 12 -

restoration path or the normal path is being used for routing.
Then, there is a list 64,70 of component identifiers, one for
each level in the hierarchy below the level of the current
label. It is to be understood the order of those fields is not
5   essential. There is a one-to-one correspondence between the
component identifiers and the components in the table described
above. In one embodiment, there may be a one bit place-holder
in the components label for each component identifier which is
set to indicate that component. More generally, a multi-bit
10   component identifier can be employed for each level. The
component identifier must contain enough bits to uniquely
distinguish between all possible components.

The actions performed at each node in
routing/switching packets will now be described in detail with
15   reference to the flowchart of Figure 7.

Step 7-1: the node obtains (for example pushes) the
LSP label (in current label field);

Step 7-2: the node obtains (for example pushes) the
components label (hierarchy);

20   Step 7-3: the node looks up the record in the table
for the LSP label;

Next, before step 7-6, one of steps 7-4, 7-5 may be
executed. Step 7-4 is executed in the event that it is time to
label switch to a higher level in the hierarchy. Neither of
25   steps 7-4 or 7-5 is executed in the event there is no need to
perform a label switch to a different level in the hierarchy.
Step 7-5 is executed in the event that it is time to label
switch to a lower level in the hierarchy.

Step 7-4: In the event the current node is the source
30   node of an LSP/FA-LSP of which the current LSP/FA-LSP forms a
component, then it is time to do a label switch. The LSP
label is switched to contain the label of the next higher level

- 13 -

LSP/FA-LSP which is used by the current LSP/FA-LSP.   The
components label is changed to include in the first component
identifier an identifier of the current LSP/FA-LSP.

5   In the event the current node is a transit node
associated with the LSP label, then there is no need to change
the first label or the hierarchy and thus neither of steps 7-4
or 7-5 is required.

Step 7-5: In the event the current node is the
destination node of the current LSP/FA-LSP, the LSP label
10   (hierarchy) is changed to the LSP label for the lower level in
the hierarchy as determined from the components label and the
table.   The components label is also changed so as to no longer
include a components identifier in respect of the lower level.

Step 7-6: The node re-applies (for example pops) the
15   components label.

Step 7-7: The node re-apples (for example pops) the
LSP label.

In all cases, the output interface would be changed
such that the packet is forwarded to the appropriate next node.

20   Now, the format of cells for our example scenario
will be described, first for the case where the normal path
between nodes 35 and 37 is used, and second for the case where
the protection path between nodes 35 and 37 is used.

Referring now to Figure 5A, shown is the cell format
25   used between nodes 31 and 33.   The cell includes the user
packet and which includes the packet, a "0" bit indicating that
the primary path is being used, and in a first label field, the
LSP label A, which is the label for the highest LSP/FA-LSP
hierarchy path.

30   Referring now to Figure 5B, shown is the cell format
used between nodes 33 and 35.   Once again, the cell includes
the user packet.   Shown in the first label field is LSP label K

- 14 -

which is the LSP label for the FA-LSP defined between nodes 33
and 38 which is being used for the current transmission. The
"0" indicates that the primary path is being used. Finally,
the components label has a single entry for indicating the

5   component identifier of the lowest level in the hierarchy, in
this case LSP A. By indicating "1st component", this means that
the label for A can be recovered by looking at the first
component identified in the table for the current LSP/FA-LSP,
namely FA-LSP K. In the event the connection was that

10   originating at end user terminal T7 through node 32, then the
lowest level LSP/FA-LSP would be LSP B the component identifier
would indicate "2nd component" from which the label for B can be
recovered by looking at the second component identified in the
table for the current LSP/FA-LSP, namely FA-LSP K.

15   Referring now to Figure 5C, shown is the cell format
for use between nodes 35 and 36. Once again, the cell includes
the user packet being transmitted. The LSP label field is
filled with "X" which is the FA-LSP from nodes 35 to 37 used
for transmission between nodes 35 and 36. The "0" again

20   indicates that the primary path is being used. Next, there are
two entries for indicating the LSP/FA-LSP of the two lower
levels in the hierarchy. The first entry indicates the
component for the next lowest level in the hierarchy used for
the path, namely component K. K is the first component of FA-

25   LSP X, and as such the entry is used to indicate the first
component. Similarly, the second entry indicates the component
for the lowest level in the hierarchy, namely component A which
was the first component of component K. In any case, the
component identifiers of lower levels in the hierarchy will

30   carry over from lower levels.

        Figure 5D shows the format used between nodes 36 and
37. This format is identical to that used between cells 35 and
36 because transmission is still within nodes belonging only to

FA-LSP X and at the same level of hierarchy. The output
interface would be changed however.

Figure 5E shows the cell format used between nodes 37
and 38. This cell format is the same as that used between
5   nodes 33 and 35.

Figure 5F shows the cell format used between nodes 38
and 14. This cell format is the same as that used between
nodes 31 and 33.

Now, in the event the restoration path for FA-LSP X
10   is activated, then the cell format is slightly changed for some
of the transmissions. The cell format between nodes 31 and 33
and between nodes 33 and 35 is unchanged from that introduced
for the normal example and is shown in Figures 6A,6B
respectively.

15   There is then a cell format between nodes 35 and 42
which is indicated in Figure 6C. In this case, the LSP label
is "X" which is the FA-LSP defined between nodes 35 and 37.
However, in this case, the restoration field is set to "1"
indicating that the restoration path is being used, namely the
20   restoration path 35,42,43,37 defined in the table for FA-LSP X.
The two component identifiers are as before, filled in to point
to LSP/FA-LSPs K and A by containing "first component", "first
component".

Figure 6D shows the cell format used between nodes 42
25   and 43 for the restoration path example. The cell format is
the same as that used between nodes 35 and 42. The output
interface would be changed such that cells are forwarded to
node 43.

Figure 6E shows the cell format used between nodes 43
30   and 37 for the restoration path example. The cell format is
the same as that used between nodes 35 and 42 although the

- 16 -

output interface would be changed such that cells are forwarded
to node 43.

Figure 6F shows the cell format used between nodes 37
and 38, for the restoration path example, this being identical
5   to the cell format used for normal example. The fact that the
restoration path was used for the higher level in the hierarchy
is hidden in this view, and no longer relevant.

Finally, Figure 6G shows the cell format used between
nodes 38 and 40 for the restoration path example, this being
10   identical to the cell format used for the normal example.

Comparison

A comparison between the overhead introduced using
traditional label stacking, and the overhead introduced using
15   the techniques provided by the above described embodiment of
the invention will now be made. In an IP network if it is
assumed that the traffic will be received via Ethernet
accesses, then packets will vary in length between 0 and 1500
bytes. These packets are not uniformly distributed in size,
20   and are likely to have a trimodal distribution (much more small
packets than large packets, with 3 important peaks).

In this comparison it is assumed that there are five
levels of hierarchy. In this example it is assumed that the
necessary hierarchy information (pointers or components
25   identifiers) could be transported in one single label. The
following table summarizes the percentage occupied by the
header when a regular label stack is employed (that described
in the Background of the Invention), and the label approach
provided by an embodiment of the invention for packet lengths
30   from 40 to 1500 bytes. It can be seen that for short packets
(40 bytes in length), the header percentage is reduced from

33.3% to 16.7% which is a very significant decrease in systems
with high short packet frequency.

Packet length in bytes	40	100	480	1000	1500
Packet length with regular label stack	60	120	500	1020	1520
Packet length with components id	48	108	488	1008	1508
Header with regular label stack (%)	33.3	16.7	4	2	1.3
Header with one label (%)	16.7	7.4	1.6	0.8	0.5

5

It can be seen that the most benefit from the method will be
realized for traffic which contains a high percentage of small
packets, for example voice traffic.

Numerous modifications and variations of the present

10   invention are possible in light of the above teachings. It is
therefore to be understood that within the scope of the
appended claims, the invention may be practised otherwise than
as specifically described herein.